Cluster Sets for the PLS Search and Compare Tool

Alan F. Karr, RTI International

3/14/2018

# Summary

This document describes the method for, and results of, clustering PLS data. The purpose of the clustering is to allow users of the PLS Search and Compare tool to compare a selected library to data-defined peers, with respect to four sets of characteristics.

# Dataset and Clustering Variables

The clustering was performed on a restricted use version of the PLS 2016 dataset provided to RTI by IMLS. Four sets of twenty clusters were created, each one defined by a set of related, “thematic” variables. The four cluster themes are:

* Service, comprising both the clientele and the volume of activity;
* Staff, focusing on the number of staff members of various types;
* Finance, including both revenue and expenditure;
* Collection, including both print and electronic materials.

Table 1 shows the PLS variables used to define each set of 20 clusters. In all, 23 variables are used to define the clusters.

**Table 1: Variables used to define the four sets of clusters.**

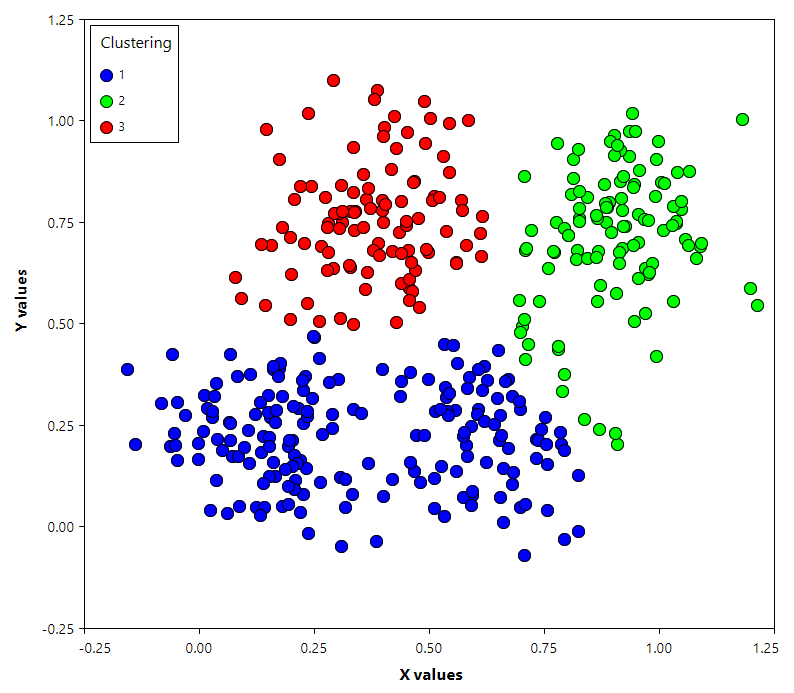
|  |  |  |
| --- | --- | --- |
| **Cluster** | **Variables** | **Number of Variables** |
| **Service** | POPU\_LSA, HRS\_OPEN, VISITS, REFERENC, REGBOR, TOTCIR, LOANTO, TOTPRO, GPTERMS | 9 |
| **Staff** | MASTER, LIBRARIA, OTHPAID | 3 |
| **Finance** | TOTINCM, STAFFEXP, TOTEXPCO, CAP\_REV, CAPITAL | 5 |
| **Collection** | BKVOL, EBOOK, AUDIO\_PH, VIDEO\_PH, ELECCOLL, SUBSCRIP | 6 |

In total, 23 variables are employed to create the clusters.

# Clustering Methodology

Cluster analysis (Hastie et al. 2016) is a statistical methodology used to group multivariate (often, high-dimensional) data on the basis of selected numerical characteristics—the “clustering variables.” It is a form of unsupervised machine learning, in the sense that clusters represent data-defined groupings rather than exogenously defined labels.[[1]](#footnote-1) The goal in forming clusters is to minimize variation of the clustering variables *within* clusters while maximizing variation of the clustering variables *across* clusters.

Figure 1 illustrates for the case of two clustering variables and three clusters.



**Figure 1: Clustering with two variables.**

The distance between a pair of data points is the Euclidean distance calculated using the clustering variables. To symmetrize the roles of the clustering variables, prior to clustering, each is standardized to have mean 0 and variance 1.

Multiple methods exist for performing clustering, the two most widely used of which are:

* **Hierarchical clustering**. All data points start in one cluster, and splits are performed recursively, moving down the hierarchy. Splits are determined in a greedy manner—the split maximizes dissimilarity of the two new formed clusters is chosen. Various termination criteria are available, the simplest of which is to proceed until a preset number of clusters are created. The results of hierarchical clustering are usually presented in a *dendrogram*. See Figure 10 below for an example.
* **k-means clustering**. Data points are partitioned into a pre-determined number k of clusters in such a manner that each belongs to the cluster with the nearest mean,[[2]](#footnote-2) which therefore serves as a descriptor of the cluster. In effect, a Voronoi tessellation has been constructed, which can be applied to cluster additional points not in the dataset.

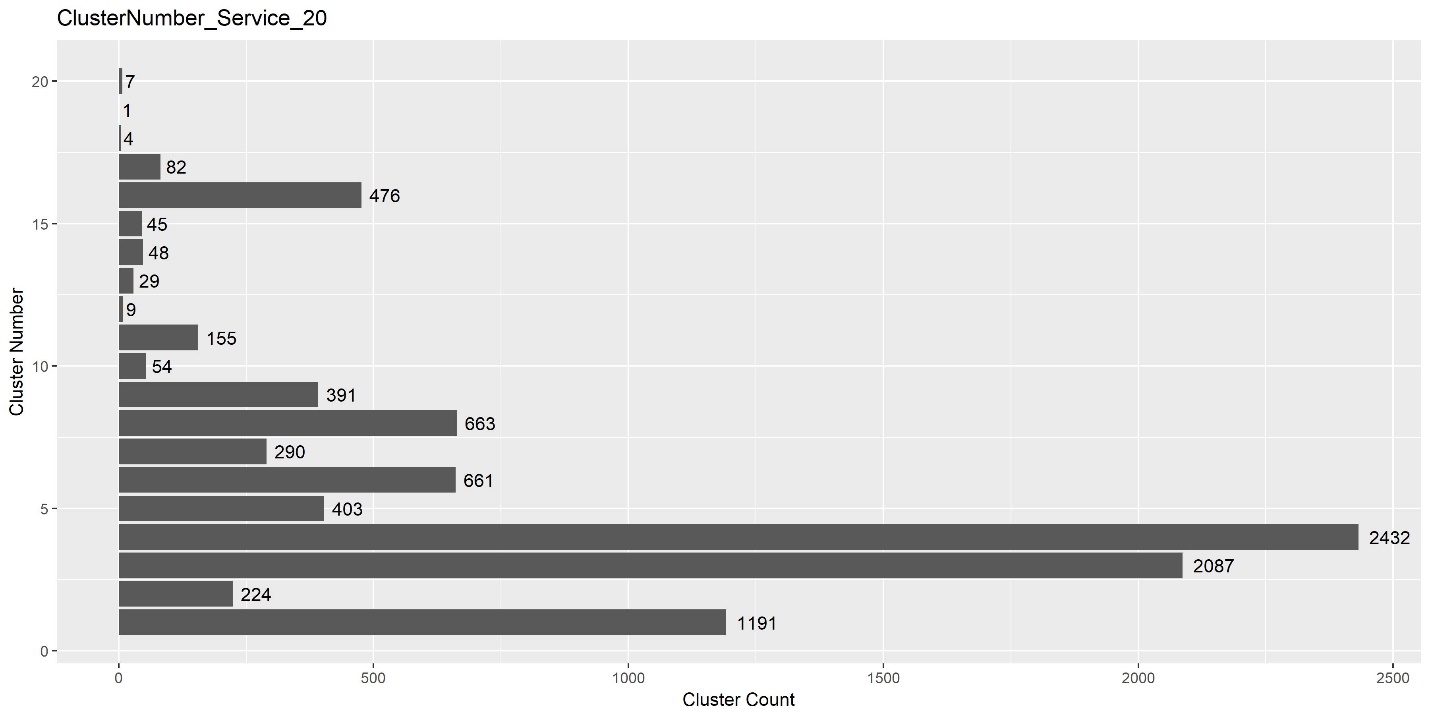
For the 2016 PLS dataset, RTI employed:

* Hierarchical clustering, as implemented in the “cluster” package of the R software environment for statistical computing and graphics (R Core Team 2017).[[3]](#footnote-3)
* The four sets of clustering variables listed above, with each standardized.
* A preset number of 20 clusters in each case. This number was chosen on the basis of empirical investigation as a compromise between the number and size of clusters. More clusters than 20 seems to yield only more small clusters while failing to split large clusters. With fewer than 20 clusters too much structure is suppressed. See Section 5 for further discussion.

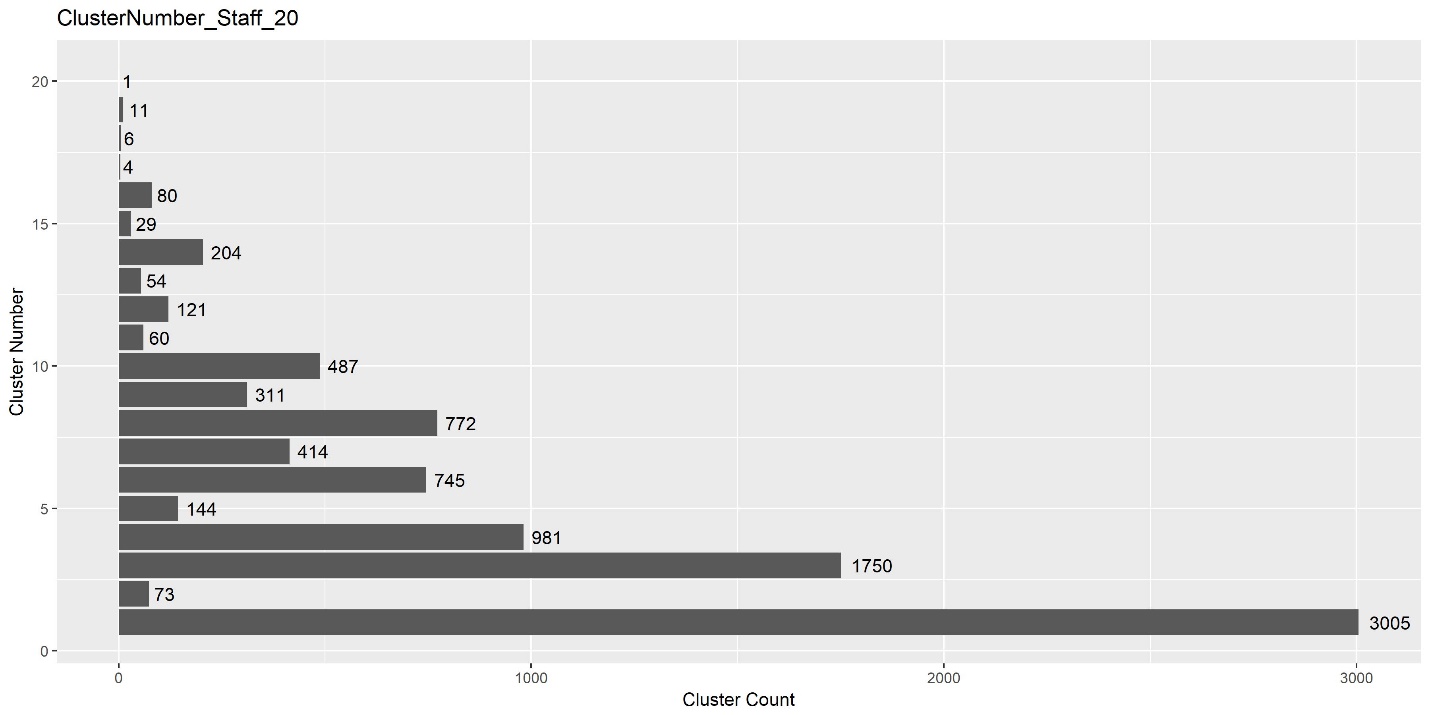
# Cluster Characteristics

Figures 2-5 show numerically and graphically the sizes of the 20 clusters within each of the four sets. The cluster numbers carry no content, and they are not comparable from one set to another. We do discuss in Section 5 the extent to which the clusters represent “independent” groupings of the data.

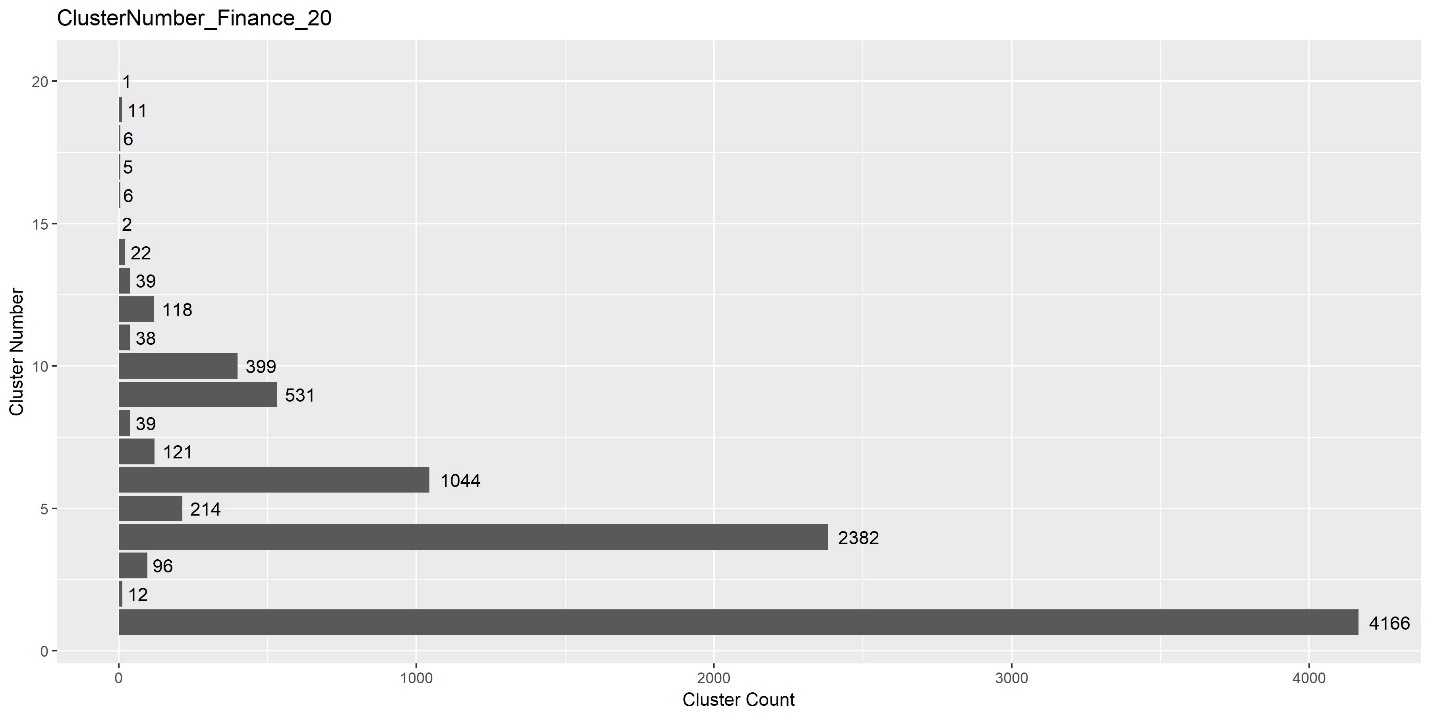
For each set, cluster sizes vary from 1 to more than 2000. While this variation is substantial, it is not alarming. From the uniformity across the four sets of clusters, it seems clear that there are large numbers of libraries that are highly similar to each other. Not surprisingly, the cluster of size 1 in all four sets consists of LIBNAME = "NEW YORK PUBLIC LIBRARY, THE BRANCH LIBRARIES."



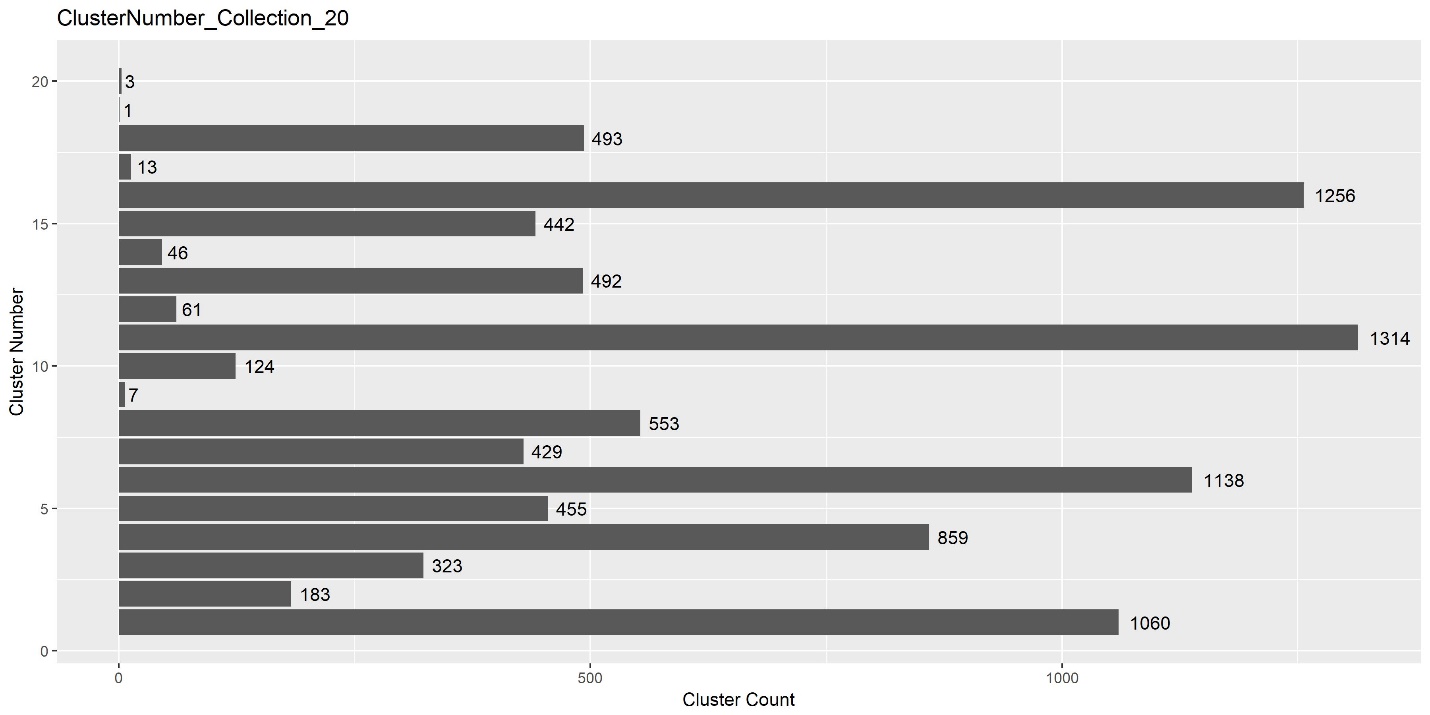
**Figure 2: Cluster sizes for the service clusters.**



**Figure 3: Cluster sizes for the staff clusters.**

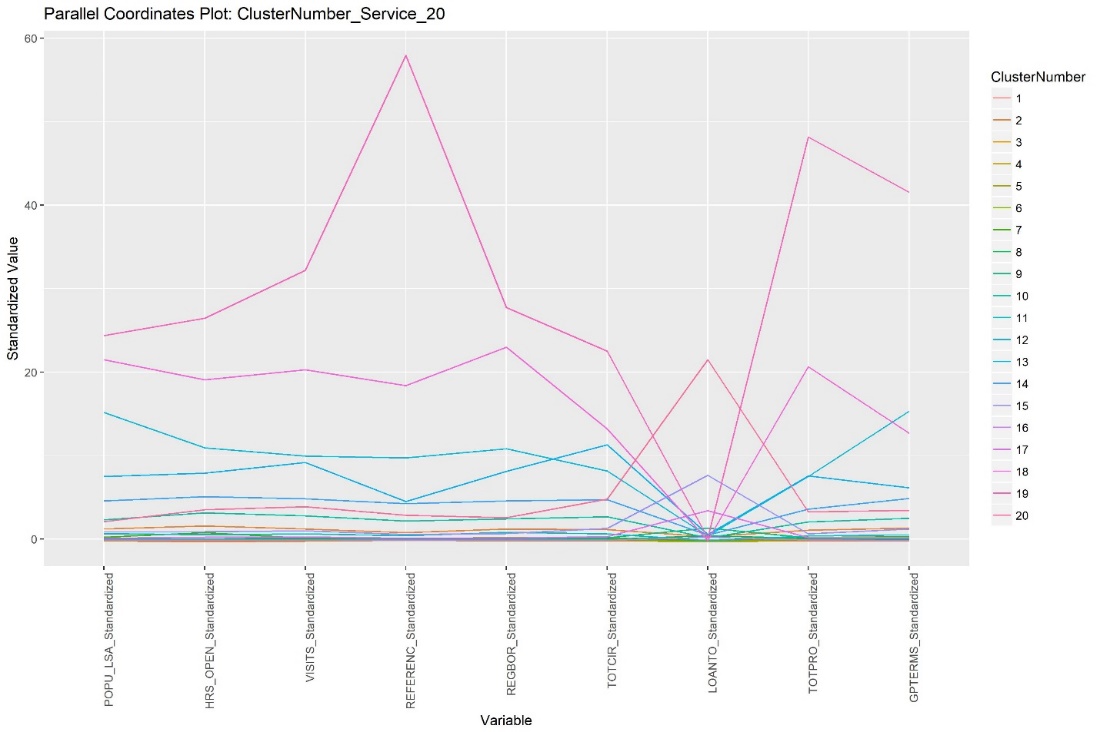


**Figure 4: Cluster sizes for the finance clusters.**

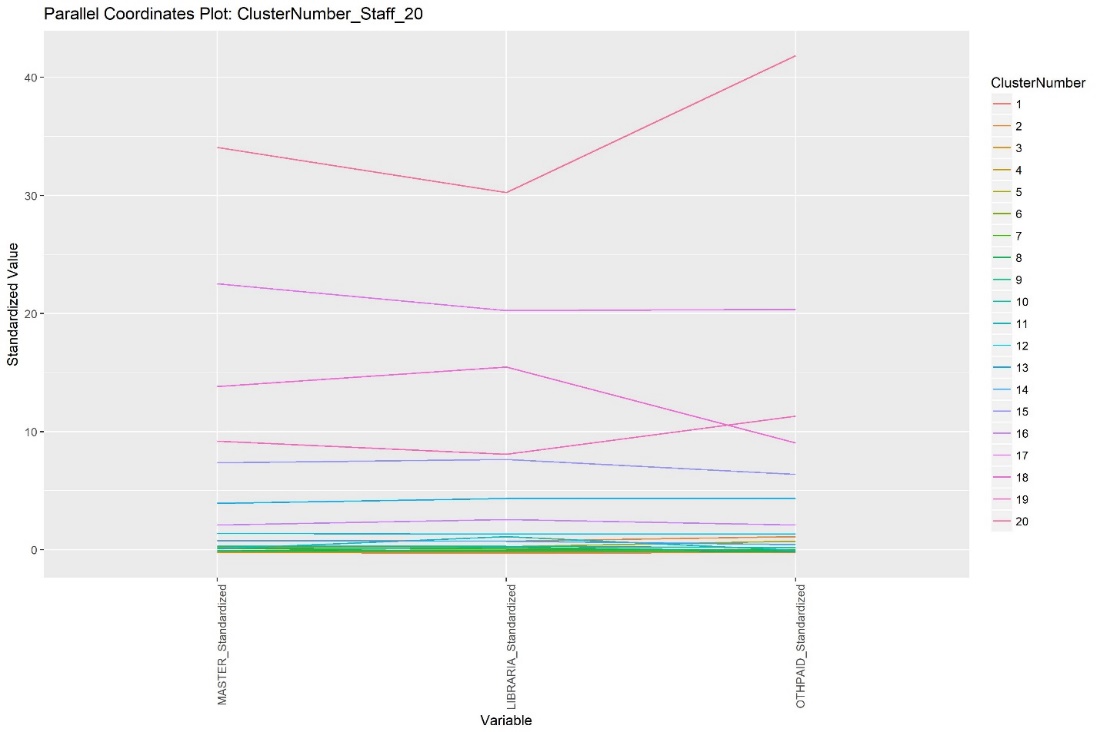


**Figure 5: Cluster sizes for the collection clusters.**

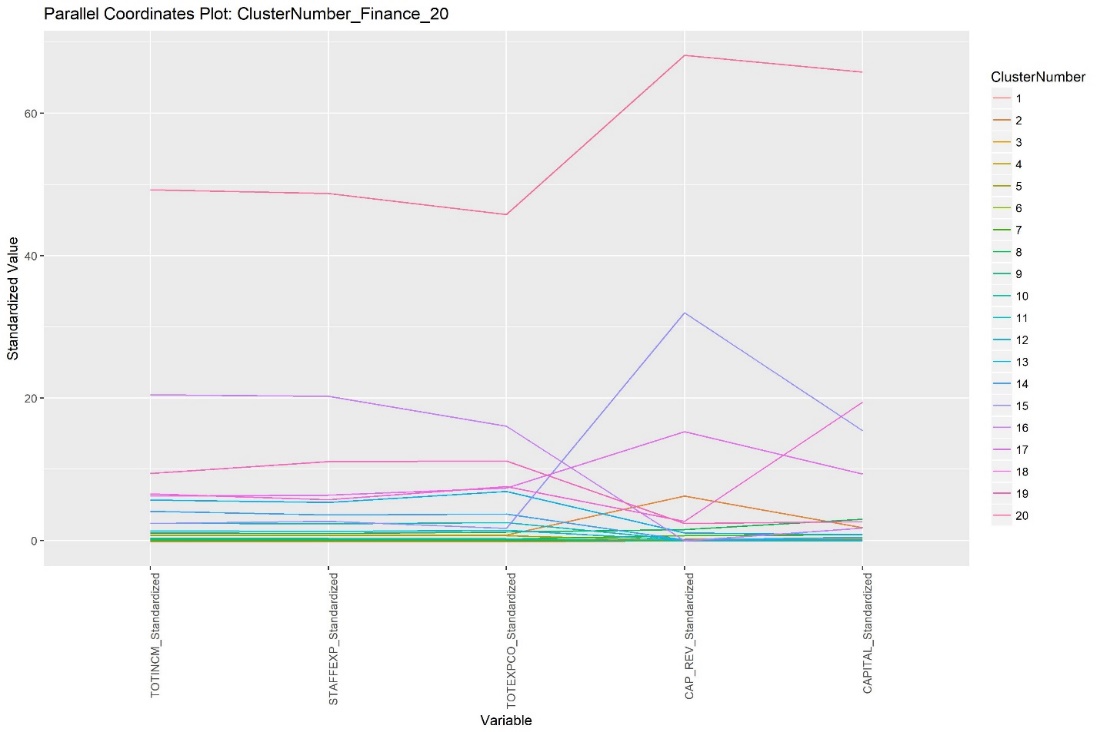
Figures 6-9 are parallel coordinates plots showing, for each of the four sets of clusters, how the standardized clustering variables behave across the clusters. In each, clusters correspond to lines (and, therefore, colors) and the mean values by cluster are plotted for each of the standardized clustering variables.



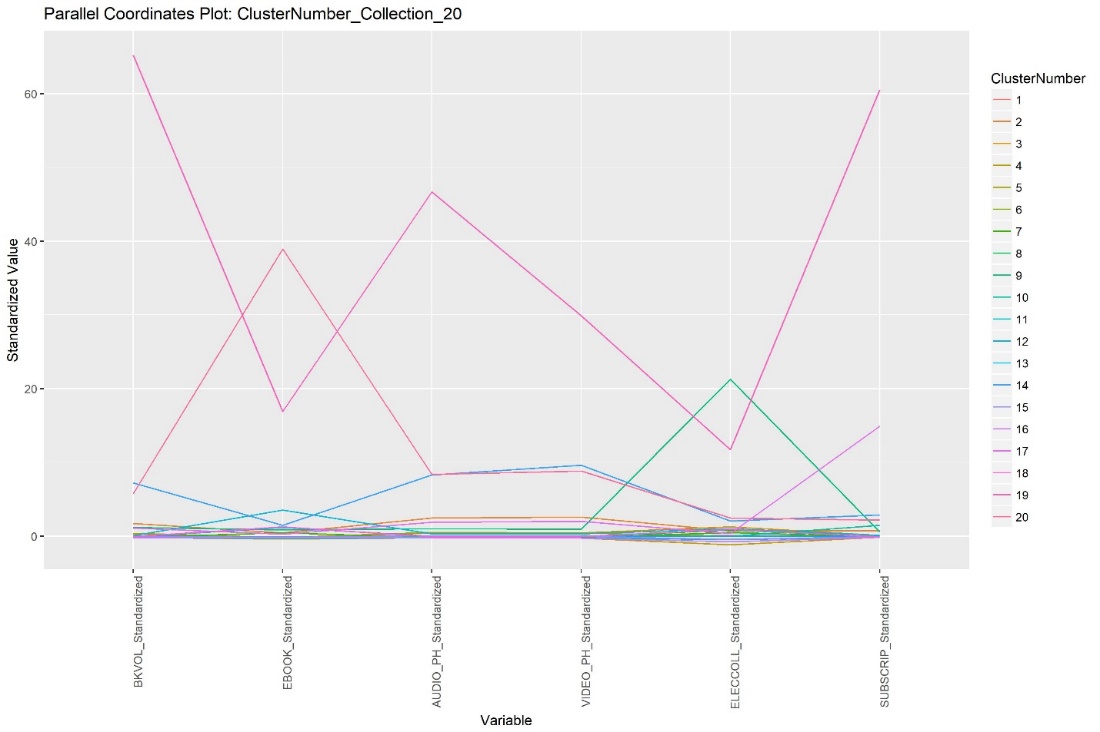
**Figure 6: Parallel coordinates plot of mean values of standardized clustering variables for the service clusters.**



**Figure 7: Parallel coordinates plot of mean values of standardized clustering variables for the staff clusters.**

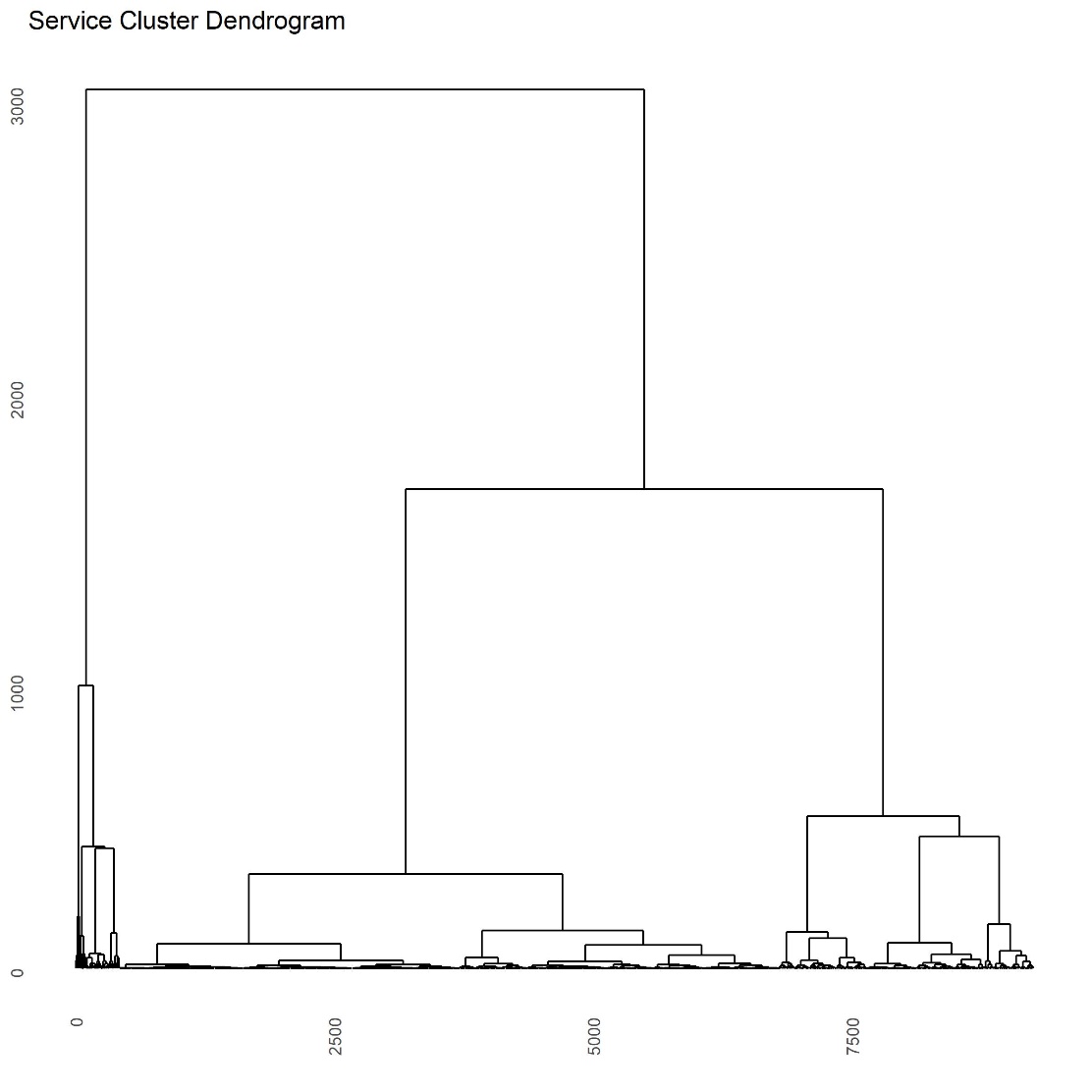


**Figure 8: Parallel coordinates plot of mean values of standardized clustering variables for the finance clusters.**



**Figure 9: Parallel coordinates plot of mean values of standardized clustering variables for the collection clusters.**

Mainly to illustrate, Figure 10 contains the dendrogram for the service clusters.



**Figure 10: Service cluster dendrogram.**

# Evaluation of the Clustering

Two questions need to be addressed:

* How “good” is the clustering?
* To what extent are the four sets of clusters “independent” groupings of public libraries?

The brief answers are “Extremely good” and “The four sets are moderately independent but are all driven by a shared ‘size’ characteristic.”

Concerning the quality of the clustering, Tables 2-5 show, for each set of clusters and for each associated, standardized clustering variable, the percentage of variation that is *explained by the clustering alone.* These values should be interpreted in the same way as correlations. All are remarkably high. We caution, however, against excessive reliance on the high values, because these result in part from the clustering’s isolating outliers such as the New York Public Library into small (at the extreme, size 1) clusters.

**Table 2: Clustering variable variation explained by clustering alone, for service clusters.**

|  |  |
| --- | --- |
| **Variable** | **Percentage of Variation Explained by Clustering** |
| POPU\_LSA | 88.4 |
| HRS\_OPEN | 87.1 |
| VISITS | 91.2 |
| REFERENC | 82.5 |
| REGBOR | 85.8 |
| TOTCIR | 84.7 |
| LOANTO | 87.6 |
| TOTPRO | 82.6 |
| GPTERMS | 85.4 |

**Table 3: Clustering variable variation explained by clustering alone, for staff clusters.**

|  |  |
| --- | --- |
| **Variable** | **Percentage of Variation Explained by Clustering** |
| MASTER | 95.1 |
| LIBRARIA | 95.2 |
| OTHPAID | 92.7 |

**Table 4: Clustering variable variation explained by clustering alone, for finance clusters.**

|  |  |
| --- | --- |
| **Variable** | **Percentage of Variation Explained by Clustering** |
| TOTINCM | 96.2 |
| STAFFEXP | 95.2 |
| TOTEXPCO | 94.0 |
| CAP\_REV | 93.7 |
| CAPITAL | 88.1 |

**Table 5: Clustering variable variation explained by clustering alone, for collection clusters.**

|  |  |
| --- | --- |
| **Variable** | **Percentage of Variation Explained by Clustering** |
| BKVOL | 82.1 |
| EBOOK | 85.6 |
| AUDIO\_PH | 78.1 |
| VIDEO\_PH | 77.1 |
| ELECCOLL | 84.0 |
| SUBSCRIP | 80.4 |

Table 6 contains the values of Cramér’s V, a measure of association for categorical variables (Cramér 1946) for the four sets of clusters. It should be interpreted similarly to correlations. The values in Table 6 constitute moderate association. Clearly, all of the 23 clustering variables are driven by some underlying, latent measure of size, so this level of association is acceptable. There is no obvious reason why collection is least associated with the other three.

**Table 6: Associations between the four sets of clusters, as measured by Cramér's V.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Service** | **Staff** | **Finance** | **Collection** |
| **Service** | 1 | .523 | .484 | .385 |
| **Staff** | .523 | 1 | .55 | .391 |
| **Finance** | .484 | .55 | 1 | .392 |
| **Collection** | .385 | .391 | .392 | 1 |

# Illustrative Application

To demonstrate how the clustering may be used with the Search and Compare tool, Table 7 shows within-cluster comparisons for a particular library, namely the Chapel Hill, NC, Public Library. The column “Value” contains the values of all of the 23 variables involved in defining any of the cluster sets for that library. The four columns under “Cluster Means” show, for each of the four cluster sets, the means of those same values for cluster that contains the Chapel Hill Public Library. The columns under Percentile Ranks show the percentile rank of the Chapel Hill Public Library within each of the four clusters that contains it. Clustering variables for each cluster are shaded in gray.

To illustrate, the value of POPU\_LSA for the Chapel Hill Public Library is 59,569, while the cluster means are 123,504 for the Service cluster containing the Chapel Hill Public Library, 54,199 for the Staff Cluster, 81,545 for the Finance cluster, and 99,755 for the Collection cluster. (The other libraries in the cluster depend on the cluster set, i.e., Service, Staff, Finance or Collection.) For POPU\_LSA, the percentile rank of the Chapel Hill Public Library within its Service cluster is 6.45, the rank within its Staff cluster is 68.17, the percentile rank within its Finance cluster is 47.12, and the percentile rank within its Collection cluster is 47.37.

**Table 7: Within-cluster comparisons for Chapel Hill (NC) Public Library.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Chapel Hill Public Library** | | **Cluster Means** | | | | **Percentile Ranks** | | | |
| **Variable** | **Value** | **Service** | **Staff** | **Finance** | **Collection** | **Service** | **Staff** | **Finance** | **Collection** |
| **POPU\_LSA** | 59,569 | 123,504 | 54,199 | 81,545 | 99,755 | 6.45 | 68.17 | 47.12 | 47.37 |
| **HRS\_OPEN** | 3,233 | 8,362 | 4,986 | 7,067 | 8,829 | 10.32 | 30.18 | 16.29 | 18.58 |
| **VISITS** | 657,976 | 482,501 | 263,267 | 345,253 | 406,685 | 87.1 | 99.38 | 95.24 | 86.07 |
| **REFERENC** | 33,228 | 88,985 | 42,061 | 60,457 | 83,951 | 19.35 | 51.33 | 32.08 | 37.77 |
| **REGBOR** | 39,342 | 78,093 | 30,159 | 44,624 | 53,990 | 13.55 | 76.59 | 55.89 | 54.8 |
| **TOTCIR** | 1,385,572 | 869,340 | 406,256 | 531,282 | 635,083 | 88.39 | 99.59 | 98.25 | 92.88 |
| **LOANTO** | - | 2,576 | 20,323 | 20,602 | 13,680 | 8.39 | 3.9 | 3.76 | 4.95 |
| **TOTPRO** | 1,020 | 1,322 | 938 | 1,217 | 1,341 | 39.35 | 70.84 | 51.38 | 51.08 |
| **GPTERMS** | 69 | 87 | 47 | 68 | 89 | 31.61 | 82.96 | 64.16 | 51.39 |
| **MASTER** | 10 | 11 | 8 | 9 | 11 | 45.16 | 78.03 | 56.89 | 56.04 |
| **LIBRARIA** | 10 | 15 | 10 | 12 | 15 | 32.26 | 56.06 | 38.85 | 42.72 |
| **OTHPAID** | 24 | 31 | 17 | 24 | 29 | 36.13 | 88.5 | 57.39 | 49.54 |
| **TOTINCM** | 2,964,608 | 4,286,731 | 2,480,386 | 3,290,112 | 3,697,314 | 29.03 | 74.54 | 36.84 | 48.3 |
| **STAFFEXP** | 2,045,371 | 2,707,610 | 1,626,954 | 2,127,611 | 2,387,595 | 33.55 | 79.47 | 47.37 | 51.7 |
| **TOTEXPCO** | 239,332 | 432,886 | 241,515 | 314,921 | 400,979 | 21.94 | 55.44 | 26.07 | 35.29 |
| **CAP\_REV** | - | 569,393 | 234,430 | 43,366 | 161,157 | 55.48 | 61.81 | 62.16 | 52.01 |
| **CAPITAL** | - | 319,185 | 249,606 | 86,865 | 323,367 | 40 | 39.22 | 39.35 | 42.72 |
| **BKVOL** | 176,543 | 216,763 | 135,043 | 175,886 | 216,880 | 32.9 | 82.14 | 60.65 | 49.54 |
| **EBOOK** | 61,647 | 50,528 | 63,355 | 63,647 | 27,584 | 78.06 | 74.13 | 73.68 | 92.26 |
| **AUDIO\_PH** | 12,913 | 15,588 | 9,460 | 11,936 | 14,002 | 46.45 | 81.11 | 62.66 | 57.59 |
| **VIDEO\_PH** | 10,247 | 21,393 | 12,555 | 16,189 | 17,820 | 17.42 | 41.27 | 21.8 | 27.55 |
| **ELECCOLL** | 81 | 54 | 56 | 56 | 89 | 87.74 | 80.9 | 84.71 | 50.15 |
| **SUBSCRIP** | 163 | 383 | 311 | 338 | 314 | 23.87 | 30.8 | 20.55 | 21.05 |

# References

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Hastie, T., Friedman, J., & Tibshirani, R. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second Edition. Springer-Verlag, New York.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

1. This latter problem is generally termed classification. [↑](#footnote-ref-1)
2. Both are multi-dimensional, “nearest” is with respect to Euclidean distance. [↑](#footnote-ref-2)
3. Virtually all statistical software packages perform clustering. [↑](#footnote-ref-3)